May 19, 2014

Errata - Logistic Regression Models

Joseph Hilbe, *Arizona State University*

Logistic Regression Models
Joseph M Hilbe

ERRATA and COMMENTS
4^{th} Printing (Printed Sept, 2010)
(updated to: 19 May, 2014)

The 4th printing enhances Stata code to use version 11 rather than version 9-10 code. The book was completed before Stata version 11 was published. For example, when constructing synthetic data, the book now uses the new Stata pseudo-random number generators rather than the ones I created back in 1995 – the suite of *rnd** commands -- or Roberto Gutierrez's unpublished *genbinomial* command.

No more corrections to the text are planned for future printings. A second edition is planned to be published in late 2014 and will include nested logistic regression, and chapters on latent class models and on Bayesian logistic models. Both single and multilevel models will be examined. Certain areas of the present edition will be re-written to assist in clarity. Any suggestions you have, or typos/errors you discover in the present printing of the first edition, will be most appreciated.

Instructors may request a gratis 187-page *Solutions Manual for Logistic Regression Models*, Chapman & Hall/CRC, ISBM: 978-1-4398-2066-7. Contact author for details (hilbe@asu.edu or jhilbe@aol.com). It is available from publisher, but I will need to give you added information.

NOTE: 4th Printing is found on page opposite the table of contents. The numbers on the line under "Printed in the United States of America..." end with the number 4 -- the last number is the printing.
Thanks to Zhehui Luo of the Michigan State Dept. of Epidemiology and students in my courses on Logistic Regression and Advanced Logistic Regression for identifying remaining typos & errors. Reginald Jordon is to be especially acknowledged for identifying several items that had not been caught for 3 years. I have added comments and additions to the actual errata. The Comments section follows the Errata, beginning with page 3.

Page xvii: Final full paragraph at the bottom of the page. The books web site should now read:
http://works.bepress.com/joseph_hilbe/

Page 1(bottom) page 2 (top): Starting from the sentence beginning with "First, the error term..." on the bottom line of Page 1, amend to read:
"First, the error terms are non-normally distributed. Second, the..."

Page 18: The terms A*D/B*C near the bottom of the page: Change to small letters to read as:
"The odds ratio is calculated by (a*d)/(b*c) or (a/c)/(b/d)."

Page 19: Near top of page: add "nolog" to first Stata command line under "MAXIMUM LIKELIHOOD LOGISTIC COMMAND". Read as:

```
". logistic death anterior, nolog"
```

Page 20: output at top of page. First BIC should be AIC.

Page 30,31: The comments to the right of the calculations of probabilities for each of the three non-reference Killip level. Delete the ending phrase for each, "with respect to KK1".

Page 39: The top Stata output in mid page: the term "ons" should read "_cons"

Page 73: The paragraph following the list of predictors in mid page. Replace the paragraph with the following paragraph:

"We would ordinarily prefer to model *age* a as continuous predictor. But this assumes that the odds ratio is the same across all ages. If we suspect that this is not the case, it is preferable to categorize the predictor into units reflective of changes in odds ratios. *Age* is indeed skewed to the right, so we shall model *age* using the four levels of *age* which have previously been made into indicator or dummy (0,1), variables. I show on the next page how the indicator variables were created, and why."

Page 86: Stata code near bottom of page: *ant]erior* should be *anterior*

Page 110 Eq 5.19 Close parentheses for both numerator an denominator.

Page 118: ">" sign between RR and left side equation should be ">"

Pages 120 and 128: The "///" symbols should be "//".

Page 130 third word, "percent", of the first full paragraph is misspelled. The sentence should read as: "The 95% confidence interval of the attributable risk is given as"

Also, same page, first line following Eq 5.39, the words, "lower" and "upper" should be reversed.

Page 132 third/fourth line under equation 5.40. Change sentence beginning with "Scaling replaces" to read as:
"... Scaling replaces *W* by the product of the model standard error and square root of the Pearson dispersion statistic."
Thus, scaled SE = se(βs) = se(β)*sqrt(Pdispersion).

Page 133: Close space between *rbinomial* and *(d,exb).*

```
CREATE BINARY LOGISTIC RESPONSE WITH DEFINED DATA; BINOMIAL DENOMINATOR=100
. gen d = 100
. gen exb = 1/(1+exp(-xb))
. gen y = rbinomial(d, exb)
```

Page 172: R code: 2nd block of code from the top of page. Should read as:

```
age1 <- ifelse(agegrp=='=<60', 1, 0)
age2 <- ifelse(agegrp=='61-70', 1, 0)
age3 <- ifelse(agegrp=='71-80', 1, 0)
age4 <- ifelse(agegrp=='>80', 1, 0)
```

Place the above code above final block of code on page 171.

page 191: Stata code near bottom of page: Replace "of" with "if" to read:

```
. drop if class==4
```

page 212: mid page. The command, " gen byte whlo = white*los" should be typed instead:
. gen whlo = white*los

Since one of the values of *los* is greater than 100, *los* must be stored as an ***integer***. ***bytes*** can range from -127 to 100. The result was the loss of the observation with los=116, and a slight change in coefficient and SE values. Page 218 has model done correctly.
Note that ***integer*** storage types range from -32,767 to 32,740. ***Long*** types range from -2,147,483,647 to 2,147,483,620. For numbers with decimals, stored as ***floats.*** If greater than -/+ 1.7*10^38 then store as ***double***.

Page 215: The command "corr, _coef cov" is no longer used in recent version of Stata. To obtain the variance-covariance matrix now, simply type "vce" on the command line.

Page 215: Amend equations 6.11 and 6.12 so that there is a bracket on the 3rd term of each

$$\text{Variance} = (r_1\text{-}r_0)^2 * V(\beta_1) + [x(r_1\text{-}r_0)]^2 * V(\beta_3) + 2x(r_1\text{-}r_0)^2 * CV(\beta_1,\beta_3) \qquad (6.11)$$

$$\text{SE} = \text{sqrt}[(r_1\text{-}r_0)^2 * V(\beta_1) + [x(r_1\text{-}r_0)]^2 * V(\beta_3) + 2x(r_1\text{-}r_0)^2 * CV(\beta_1,\beta_3)] \qquad (6.12)$$

Page 216: Lower part of page,
First line of code (numbers) under 95% CONFIDENCE INTERVALS AT LOS=1, should read:

```
. di (.7709236 -.0477605*1) - 1.96* sqrt(.087415 + 1^2 * .000335+2*1*(-.003864))
.16871516
```

In the book the final term is missing a negation sign. It should read '-.003864'. It is correct in other places on the page.

Page 217. 7 lines from the top. The correct value for the upper confidence interval of the odds ratio is 3.5880576, not 3.26671, which is the exponentiation of the lower CI of the odds ratio. To get the correct value one exponentiates 1.277611, i.e. exp(1.277611) = 3.5880576.

Page 217: Section 6.4.5, line 5. IRR should read OR.

Page 217: The formula used to calculate a p-value near the bottom of the page is mistaken. See page 104 for explanation. The last Stata code and output on the page should read as:

```
. di (1-normprob(1.404184))*2
.16026407
```

The corresponding R code is (for pages 239/240)

```
> pnorm(1.40184, lower.tail=F)*2
```

Page 219, Figure 6.4 Stata's graph commands have changed since first written. New code is:

```
. scatter xb0 xb1 los, connect(l l) symbol(O d) xlabel(0 10 to 100) sort
   l1title(Predicted logit) title(Interaction of White and LOS)
```

Page 220, Figure 6.5 Stata's graph commands have changed since first written. New code is:

```
. scatter yhat0 yhat1 los, connect(l l) symbol(O d) xlabel(0 10 to 100)
   sort l1title(Predicted logit) title(Interaction of White and LOS)
```

Page 227: 5 lines from the bottom, first term and number in line. "80" should read "90".

.

Page 236. Section 6.1, line 3. "tpass <-" typed twice. Delete one of them.

Page 237, lines 7 and 8 from top.

1) comment should read, "man:woman give age=adult / man:woman|age=child
2) next line, "#no plotgr3 in R" should read "#no postgr3 in R"

Page 247 Section 7.1.2

Substitute the table below for the one in the book.

MODEL	DEVIANCE	DIFFERENCE	DF	MEAN DIFFERENCE
intercept	1486.290			
MAIN EFFECTS				
anterior	1457.719	28.500	1	28.500
hcabg	1453.595	4.124	1	4.124
killip	1372.575	81.020	3	27.001
agegrp	1273.652	98.923	3	32.974

Page 259. Add sentence to the end of Section 7.3, just above 7.3.1

"In general, BIC statistics give greater adjustment weight to the number of predictors in the model than does AIC. "

Page 263: Section 7.3.3-7.3.5 to be amended to read as follows. Substitute the text between the double-double lines for what is now in the book. My apologies for the inconvenience.

Delete the current section 7.3.3 LIMDEP AIC. It is appropriate for normal models, not logistic models.

===

PAGE 263 TO MID 267

===

===

7.3.3 Other AIC statistics

There have been a number of AIC-type statistics developed since Akaike first constructed his information criterion in 1973. Two others that have found considerable use are both called corrected AIC statistics. The first was by Sigiura (1978), formulated as

$$AIC_c = AIC + \frac{n(p+1)}{n-p-2}$$

(7.24)

Simulation studies have shown it to have less bias than the AIC, and to perform better than AIC when n/p is small.

The second *corrected* version was by Bozdogan (1987). He defined the equation as

$$CAIC = -2L + p\{ln(n) + 1\}$$

(7.25)

Bozdogan criticized Akaike's original formulation of AIC due to the fact that it does not depend on sample size. Because of this he showed that it lacked the properties of asymptotic consistency. Sigiura's definition addresses the same problem. Studies have demonstrated that AIC_c in particular is preferred to AIC for assessing comparative model fit. We use AIC_n instead for most of our comparative analyses. It appears to be able to select the best fitted model as well as the statistic with more terms.

7.3.4 BAYESIAN INFORMATION CRITERION (BIC)

The BIC statistic was first developed by Gideon E. Schwarz of Hebrew University, Jerusalem, in 1978. His formulation was in response to Akaike's 1973 information criterion, whereby more weight is given to the number of predictors in the model. Schwarz also included a term for sample size, which all subsequent formulations of AIC or BIC after the original AIC have done. The philosophical basis of Schwarz's information criterion is Bayesian, unlike Akaike, but the resulting AIC and BIC equations have typically differed by only a term or so. The rationale for the two types of information criteria differ, but the resulting formulae are similar. The reasons for this go beyond the scope of this book.

Schwarz's Bayesian Information Criterion (BIC), also referred to as simply *Schwarz Criterion* in SAS output, is given as

$$BIC_S = -2LL + k*ln(n) \qquad (7.26)$$

The model with a lower BIC statistic is regarded as the better fitted model. The models being compared may be nested, but need not be. The models may be of different sample sizes as well. Comparisons of treatment and control data are common applications of the BIC statistic.

For an example, we model the same predictors as before, but use the GLM logistic model. I follow estimation with a command called *abic*, which produces two AIC statistics and two BIC statistics. The AIC-BIC pair in the right column are the statistics that are commonly displayed following Stata maximum likelihood estimation. The Stata command, *estat ic*, displays these statistics to the screen.

```
. glm death anterior hcabg kk2-kk4 age3 age4, nolog fam(bin) eform

Generalized linear models                          No. of obs      =      4503
Optimization     : ML                              Residual df     =      4495
                                                   Scale parameter =         1
Deviance         =  1276.319134                    (1/df) Deviance =   .283942
Pearson          =  4212.631591                    (1/df) Pearson  =  .9371817
```

```
Variance function: V(u) = u*(1-u)                    [Bernoulli]
Link function    : g(u) = ln(u/(1-u))                [Logit]

                                                     AIC             =  .2869907
Log likelihood   = -638.1595669                      BIC             = -36537.86
------------------------------------------------------------------------------
             |                 OIM
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    anterior |   1.894096   .3173418     3.81   0.000     1.363922    2.630356
       hcabg |   2.195519   .7744623     2.23   0.026     1.099711    4.383246
         kk2 |   2.281692   .4117012     4.57   0.000     1.602024    3.249714
         kk3 |   2.218199   .5971764     2.96   0.003     1.308708    3.759743
         kk4 |   14.63984   5.218374     7.53   0.000     7.279897    29.44064
        age3 |   3.549577   .7119235     6.32   0.000     2.395823    5.258942
        age4 |   6.964847    1.44901     9.33   0.000     4.632573    10.47131
------------------------------------------------------------------------------

. abic
AIC Statistic =   .2869907            AIC*n      = 1292.3191
BIC Statistic =   .2908262            BIC(Stata) = 1343.6191
```

Notice that the model output includes an AIC and BIC statistic. The displayed BIC statistic is -36537.86, which is different from Stata's BIC which is displayed in the *abic* output --- with a value of 1343.6191. The value of AIC in the model output is the same as indicated by (left column) AIC in the *abic* output. The model BIC is based on the Deviance definition, and is expressed as

$$BIC_R = D - df * \ln(n) \tag{7.27}$$

We may calculate it as follows,

```
di 1276.319134 - 4495*ln(4503)   // LL - dof*ln(n)
-36537.864
```

which gives the same value as shown in the above model output. This version of BIC was specifically designed by the University of Washington's Adrian Raftery in 1986 to be used with GLM software. GLM algorithms generally base model convergence on the deviance function, and a few applications do not even estimate a log-likelihood function during the modeling process. The BIC_R statistic is given rather than Schwarz's statistics due to the GLM estimation environment. *abic* provides Schwarz's criterion, giving us both versions to be used for comparative analysis.

The BIC_S statistic may be calculated as

```
. di -2*(-638.15957) + 8 * ln(4503)
1343.6191
```

which is identical to the value displayed in *abic* results. The *abic* right column statistics can be produced using Stata's *estat ic* command.

```
. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |   4503           .   -638.1596      8     1292.319    1343.619
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Raftery developed a table providing the degree of model preference based on the absolute difference between the BIC statistics of two models.

|difference|	Degree of preference
0 - 2	Weak
2 - 6	Positive
6 - 10	Strong
> 10	Very Strong

Models A and B:
If $\text{BIC}_A - \text{BIC}_B < 0$, then A preferred
If $\text{BIC}_A - \text{BIC}_B > 0$, then B preferred
or
Model with lower BIC value preferred.

For the example models we have worked with in this chapter, the reduced model has the following partial output.

```
                                                     No. of obs      =      4503
                                                     Residual df     =      4497
                                                     AIC             =  .3074784
Log likelihood   = -686.2875063                      BIC             = -36458.43

. abic
AIC Statistic =   .3074783            AIC*n      = 1384.5751
BIC Statistic =   .3095883            BIC(Stata) = 1423.05
```

The deviance based BIC statistics are : -36537.86 to -36458.43
The log-likelihood based BIC statistics are: 1343.62 to 1423.05

DEVIANCE

```
. di -36537.86 -(-36458.43)
-79.43
```

LOG-LIKELIHOOD

```
. di 1343.62  -  1423.05
-79.43
```

Both differences are identical. This relationship maintains for other nested models as well. In either case, however, the absolute difference between the full and reduced model is substantially greater than 10, indicating a very strong preference for the full model.

The AIC and BIC statistics give us consistent advice. Both the AIC and BIC tests tell us that the full model is preferred.

Recall that the true value of the AIC and BIC statistics rests in the fact that they can both compare non-nested models. For example, modeling the same data using the full model with a Bernoulli *loglog* link, the likelihood BIC statistic is 1335.4843. Compare this to the same statistic *logit* link value of 1343.6191. The difference is 8.13, in favor of the *loglog* link. This indicates that the *loglog* link is strongly preferred over the *logit* link.

Note that the deviance statistic is used only for GLM-based statistical procedures. The log-likelihood is the normal way to estimate all other maximum likelihood models. Most contemporary GLM algorithms have, though, a calculated log-likelihood function as part of the output, therefore the deviance-based formula is rarely used now.

7.3.5 HQIC GOODNESS OF FIT STATISTIC

The HQIC, or Hannan and Quinn Information Criterion (Hannan & Quinn,1979), is defined as

$$\text{HQIC} = -2\{LL - k*\ln(k))/n \qquad (7.28)$$

The calculated values for the nested models we have been discussing in this chapter are:

```
. di -2*(-638.15957 -8*ln(8))/4503
.29082616

. di -2*(-686.28751 -6*ln(6))/4503
.3095883
```

The HQIC test is an alternate version of BIC, used in LIMDEP. If the values of the BIC and HQIC differ greatly, it is wise to check the models.

It is vital to be certain you know which version of AIC and BIC is being used. Whatever the version, be consistent throughout the comparative evaluation of models.

7.3.6 A Unified AIC Fit Statistic

==

==

==

Page 272. Equation 7.31, Parentheses are needed for the denominator of the second term within brackets, y/(m*μ)). The equation should read as:

$$d = +/- \text{sqrt}[2\Sigma y * \ln(y/(m*\mu)) - (m - y)*\ln((m - y)/(m*(1-\mu))]$$

Page 293. R code: 3rd line from top.

Use: library(PresenceAbsence)

in place of: library(epicalc)

Same block of code: amend to read:

```
cmx(cmxdf, threshold=0.05, which.model=1)  #confusion matrix
```

Page 299; Equation 8.14: the final term should read $ln\binom{m}{y}$.

Following 8.16 add (not a correction - an enhancement)

or

$$D = 2\Sigma\{y\ln(1/\mu) + (m-y)\ln(1/(m-\mu))\} \qquad (8.16a)$$

Page 300: Code in mid-page. Should read as:

```
Dev= 2Σ{yln(1/μ) + (m-y)ln(1/(m-μ)) }
```

$$\chi^2 = \Sigma\ (y-\mu)^2\ /\ (\mu*(1-\mu/m))$$

$$LL = \Sigma\{\ y*\ln(\mu/m)+(m-y)*\ln(1-(\mu/m))\}$$

NOTE: The deviance in the book is OK, but this has better convergence properties. See Hilbe & Robinson, *Methods of Statistical Model Estimation* (2012), Chapman & Hall/CRC.

Page 302 line 5 of the text. The sentence should read:
"... freedom is the number of observations in the model minus the number of predictors, including the intercept. For a model parameterized as odds ratios, an exponentiated intercept is assumed, if not displayed, in the output. However, such an intercept has little useful interpretation".

Page 322: Amend

```
 . save overex  /// save data
```

to

```
. save overex  // save data
```

Page 323: Delete "/// a user authorized command" near the top right of the page.

Page 335: Delete the "[" at the start and "]" at the end of the long line of Stata code in middle of page.

Page 350: Code was dropped from the book in printing that was in my manuscript -- the poissonX2 function. Below I have included the function as it should exist in the book.

```
fit9_2i<- glm(studytim ~ drug, data=cancer, family=poisson) 131
poissonX2 = function(y, e) { #Compute Pearson chi-square for poisson
#y: number successes
#e: expected probability
return(sum((y-e)^2/e))
}
summary(fit9_2i)
fit9_2iX2<- poissonX2(cancer$studytim, fitted(fit9_2i, type='response'))
cat('Pearson X2:', fit9_2iX2, 'Dispersion:', fit9_2iX2/fit9_2i$df.residual,
'AIC/d.f.:', fit9_2i$aic/fit9_2i$df.residual, '\n')
```

Note: For *PearsonX2.r* function information, see 'Comments' section of this document.

Page 351, near bottom: "id <- ln(d)" should read "id <- log(simul$d)"

Page 357: The denominator of "Category or Level 3" should have all negative signs, not the two positive signs. The formula should read

$$\text{Logit} = \ln[(p_1 + p_2 + p_3)/\ (1 -\ p_1 - p_2 - p_3)]$$

Page 368: Box 10.1: the section on *white* should read as
white: The expected odds of being admitted to the hospital as an emergency patient is some 40 % less among those who identified themselves as white compared with those who identified themselves as non-white, holding the other predictors constant

hmo: Patients belonging to an HMO…

Page 376: Line immediately above Section 10.4: change words "a higher level" to "*Emergency*".

Page 387: The first word, “The”, of the paragraph immediately under equation 11.9 is mistaken. The paragraph should start out as:
“It is important to remember that the above parameterization is based on set-“

Page 388: the table about ¼ a page from the top has the 0 and 1 values in the wrong places. It should instead read as:

		Response	
		0	1
Predictor	0	A	B
	1	C	D

If you find additional errata, please advise. I will post them to this Errata page in the future. Thank you to those who have identified typos. I will list your names in the second edition.

Page 518: Section 13.4. Use the **lme4** package with the *glmer* function rather than *nlme4*. Newer packages exist for estimating these types of models; e.g., **glmmADMB**. Also see the **sabre** package.

Pages 546-547: We use the grouped 8 observation **hiv** data for page 546, but switch to the **hiv1** data for the model at the bottom of page 547, which is an observation-based models with 47 patients. The bottom line of texts on page 547 can be amended to read
“We now model *hiv* on levels of *cd4* and *cd8* using the observation-based **hiv1** data.“

REFERENCES

p 621: Replace the current reference for Swartz, J to read as:
Schwarz, G (1978). Estimating the dimension of a model, *Annals of Statistics,* Vol 6, 2:461-464.

Add:
Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika,* 52, 345-370.

Sugiura, N (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics.* A 7, 13-26.

COMMENTS

Page 65: I probably should have added the formula for the second derivative of the Bernoulli link function under Equation 4.12.

$$g''(\mu) = (2\mu - 1)\left(\frac{\partial\eta}{\partial\mu}\right)^2 = \frac{(2\mu - 1)}{\mu^2(1-\mu)^2}$$

(4.12a)

Page 67: Table 4.1 provides a schematic algorithm for the estimation of a binary logistic model. I have provided full working code for estimating a generic logistic regression using Stata and R. I display the code and output below for each below the final Comment in this section.

Page 132 Comment: In R, the *glm quasibinomial* family is the same as scaling the binomial logistic model standard errors by the Pearson dispersion statistic. I recently discovered that the R *vcov()* function that is used by programmers for calculating standard errors in fact creates scaled standard errors. This results in the SEs of models using *sqrt(diag(vcov(modelname)))* for calculating SEs to have different SEs from Stata, SAS, and other applications, particularly when the data is correlated. Dividing the displayed SEs by the square root of the Pearson dispersion statistic produces model SEs. R's *glm()* function, which is used for estimating both binary and grouped logistic models, adjusts SEs so that model SEs are displayed in the results.

. Page. 198: **xi3** and **postgr3** are used for a graph at the bottom of the page. The explanation of these commands are given a bit later on pages 199-201.

Page 300 Suggestion: the deviance function as presented is the standard one shown in texts. However, it does not work properly if used in an R GLM program. A much more simple and suitable expression for the equation, requiring less memory, is the following:

```
 Dev = 2Σ{y*ln(1/μ) + (m-y)*ln(1/(m-μ))}
```

Page 350/351: PearsonX2.r has been changed to P__disp.r. It can be used as a function in the **msme** package when loaded. The code is:

```
==========================================================
# Function to calculate Pearson and Pearson dispersion
#   following glm and glm.nb:  source(P__disp.r)
# x=modelname: ex: P__disp(mymodel)  30Jan,2012 J. Hilbe
P__disp <- function(x) {
   pr <- sum(residuals(x, type="pearson")^2)
   dispersion <- pr/x$df.residual
   cat("\n Pearson Chi2 = ", pr ,
       "\n Dispersion   = ", dispersion, "\n")
}
==========================================================
```

Joseph M Hilbe: hilbe@asu.edu or jhilbe@aol.com

STATA USER AUTHORED LOGIT COMMAND. First published in the November 2005 issue of *The American Statistician* in a review of Stata. The review may be obtained from my BePress Selected Works site, http://works.bepress.com/joseph_hilbe/

```
============================================================
*! version 1: LOGISTIC REGRESSION   :IRLS METHOD OF  ESTIMATION
* Joseph Hilbe: TAS - Stata 9.0 review: 7Jul2005
program define  jhlogit
version 6
set type double
syntax  varlist(default=none) [if] [, EForm]
```

```
gettoken y varlist : varlist
if `"`if'"' != `""' {
        preserve                /* ensure the dataset returns at end of pgm */
        keep `if'               /* retain only estimation sample */
}
if "`eform'" !=  "" { local eform "eform(Odds Ratio)" }
qui  {
tempvar mu eta u w z dev oldev llike chi2 aic  bic
* INITIALIZATION OF MU AND ETA
count
local nobs =  _result(1)
gen `mu' = (`y' +  0.5)/2
gen `eta' = ln(`mu'/(1-`mu'))
* VARIABLE  INITIALIZATION
local i      1
gen `u'     =0
gen `w'     =0
gen `z'     =0
gen `dev'  =1
gen `oldev'=1
gen `chi2'  =1
local ddev  1
* IRLS SCORING
while (abs(`ddev')> 1e-6 ) {
replace  `u' = (`y'-`mu')/(`mu'*(1-`mu'))
replace  `w' = `mu'*(1-`mu')
replace `z' = `eta'  + `u'
regress `z' `varlist' [iw=`w'], mse1  dep(`y')
drop  `eta'
predict  `eta'
replace `mu'   =  1/(1+exp(-`eta'))
replace `oldev'=  `dev'
replace `dev'  = ln(1/`mu')  if `y'==1
replace `dev' =   ln(1/(1-`mu')) if `y'==0
replace  `dev'  = sum(`dev')
replace  `dev'  = 2*`dev'[_N]
local  ddev     = `dev' -  `oldev'
local  i        = `i'+1
}
local npred =  _result(3)               /* number of predictors */
local df    = `nobs' - `npred' -  1     /* degrees of freedom   */
* CALCULATION OF  LOG-LIKELIHOOD AND GOF STATISTICS
egen `llike' =  sum(`y'*ln(`mu')+(1-`y')*ln(1-`mu'))
gen `aic' = (-2*`llike' +  2*`npred')/`nobs'          //    AIC/observations
}
* PUT VALUES INTO MATRIX
qui regress, noheader `eform'
tempname b  V
mat `b' =  get(_b)           /*  coefficient vector */
mat `V' =  get(VCE)          /*  variance-covariance matrix */
mat post `b' `V', depname(`y')  obs(`nobs')
* OUTPUT
di " "
di in gr "Logistic  Estimates"
mat mlout, `eform'
di in gr _col(1) "Observations = " in  ye `nobs' in gr _col(53) "Deviance     =   " in ye  `dev'
di in gr _col(53) "Loglikelihood = "  in ye `llike'
di in gr  _col(1)  "AIC Statistic =  " in ye `aic'
set type double
end
================================================================================
```

USE OF COMMAND

```
. use medpar           /* dataset explained in text, Ch 5.11; p. 159 */

. jhlogit died hmo white

Logistic  Estimates
------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hmo |  -.0122465   .1489251    -0.08   0.934    -.3041342    .2796413
       white |   .3033872   .2051795     1.48   0.139    -.0987573    .7055318
       _cons |  -.9261862   .1973903    -4.69   0.000    -1.313064   -.5393082
------------------------------------------------------------------------------
Observations = 1495                                   Deviance     =   1920.602
                                                      Loglikelihood = -960.301
AIC Statistic =  1.2873592
```

R -- Bernoulli or binary logistic regression -

```
====================================================================
# BINARY LOGISTIC REGRESSION. BASIC FUNCTION   7 July, 2011
# From: Hilbe, J.M and A.P Robinson (2012), Methods of Statistical Model
#    Estimation, Chapman & Hall/CRC
irls_logit <- function(formula, data, tol=.000001) {  # irls_logit options
  mf <- model.frame(formula, data)              # define model frame as mf
  y <- model.response(mf, "numeric")            # set model response as y
  X <- model.matrix(formula, data = data)       # predictors in matrix X
  if (any(is.na(cbind(y, X)))) stop("Some data are missing.")
  mu <- (y + .5)/2                              # initialize µ
  eta <- log(mu/(1-mu))                         # initialize η
  dev <- 2 * sum( y*log(1/mu) + (1 - y)* log(1/(1-mu)) )
  deltad <- 1                                   # initialize deltad = 1
  i <- 1                                          # initialize i=1
  while (abs(deltad) > tol ) {                    # IRLS loop begin
    w <-  mu*(1-mu)                               # weight
    z <- eta + (y - mu)/w                         # working response
    mod <- lm(z ~ X-1, weights=w)                 # weighted regression
    eta <- mod$fit                                # linear predictor
    mu <- 1/(1+exp(-eta))                         # fitted value; probability
    dev.old <- dev                                # setup for convergence
    dev <- 2 * sum( y*log(1/mu) + (1 - y)* log(1/(1-mu)) ) # deviance
    deltad <- dev - dev.old                       # test of 2 iterations
    cat(i, coef(mod), deltad, "\n")               # iteration log
    i <- i + 1                                    # recalibrate iter number
  }
    beta <- mod$coef                            # save coefficients
    pr <- sum(residuals(mod, type="pearson")^2)       # calc Pearson disp
    prdisp <- pr/mod$df.residual
    return(list(coef = coef(mod),                     # coef & SE display
             se = sqrt(diag(vcov(mod)))/ sqrt(prdisp)))
}
========================================================================
```

USE -- how *source()* is defined is based on where *irls_logit.r* is stored on your computer. It will be a function in the *msme* library later in 2011 (download from CRAN).

Coefficients and model standard errors are displayed. Confidence Intervals, Z statistic, and p-values can be easily calculated. Note that the scaled SEs calculated by vcov() are amended to produce true model SEs.

NOTE: A complete description of OLS, IRLS, maximum likelihood, EM, simulation, and other major methods of estimation can be found in **Hilbe and Robinson, *Methods of Statistical Model Estimation*, Chapman & Hall/CRC**. The *irls_logit* function is fully described as an example of IRLS estimation. Other more complex IRLS models are also discussed. In addition, we created a *glm*-like function called *irls*, which corrects what we believe to be shortcomings in *glm()* and *glm.nb()*, describing its modular logic the specifics of the code. After the *msme* library is loaded, *irls()* will be able to be used like *glm()* is now, together with a *summary()* function. *irls()*, however, provides a much more extensive list of post-estimation statistics. The book was published May 28, 2013.

```
> library(COUNT)   # Package associated with my Negative Binomial Regression
> source("c://rfiles/irls_logit.r")    # locate where function is saved
> data(medpar)

> i.logit <- irls_logit(died ~ hmo + white, data=medpar)
1 -1.051936 -0.01343265 0.318181 1064.628                  # iteration log
2 -0.9224268 -0.01216259 0.3017145 -4.193304
3 -0.9261831 -0.01224641 0.3033848 -0.001737683
4 -0.9261862 -0.01224648 0.3033872 -3.808509e-10

COEFFICIENTS and STANDARD ERRORS
> i.logit
X(Intercept)         Xhmo       Xwhite
 -0.92618620  -0.01224648   0.30338724

$se
X(Intercept)         Xhmo       Xwhite
   0.1973903    0.1489251    0.2051795

LOWER 95% CONFIDENCE INTERVAL
> i.logit$coef - 1.96*i.logit$se
X(Intercept)         Xhmo       Xwhite
 -1.31346050  -0.30443326  -0.09916926

UPPER 95% CONFIDENCE INTERVAL
> i.logit$coef + 1.96*i.logit$se
X(Intercept)         Xhmo       Xwhite
  -0.5389119    0.2799403    0.7059437
```

The Z-statistic and P-values may be easily calculated from the above, but the confidence intervals will indicate if a predictor is significant as well. When odds ratios are displayed, exp(i.logit$coef), recall that the standard errors are determined using the delta method, which in this case is quite simple: exp(β)*se(β); ie.

ORse <- exp(i.logit$coef)* i.logit$se.

See page 35 in text.

Also of possible interest to readers, other books of mine which have been recently published are:

Hardin & Hilbe, ***Generalized Linear Models and Extensions, third edition*** (Stata Press--Chapman & Hall/CRC) [GLME3] was published May 23, 2012.

Hilbe, (ed) ***Astrostatistical Challenges for the New Astronomy***, Springer, was published Nov 7, 2012

Hardin & Hilbe, ***Generalized Estimating Equations, 2nd edition*** (Chapman & Hall/CRC) [GEE2] was published December 10, 2012..

Hilbe and Robinson, ***Methods of Statistical Model Estimation*** (Chapman & Hall/CRC) was published May 28, 2013,

Shults and Hilbe, ***Quasi-Least Squares Regression*** (Chapman & Hall/CRC) was published Jan 29, 2014

Zuur, Hilbe, and Ieno, ***A Beginner's Guide to Modeling GLM and GLMM: a frequentist and Bayesian perspective for ecologists*** (Highlands Statistics) was published June 10, 2013.

Hilbe, ***Modeling Count Data*** (Cambridge University Press) is completed and is due to be published in June 2014.

Miner, Bolding, Hilbe, et al, ***Practical Predictive Analytics and Decisioning Systems for Medicine*** (Elsevier) is due to be published in the late summer 2014.

Thanks to Gary Anderson of the University of Melbourne for catching a number of typos and places that needed attention. Also I thank Dongjae Jung, Fabiana MacMillan, Elmira Mukailova, Jorge Rosas, Mark Jones, Mark David, Alan Burt, Christer Johansson, Rhondda Jones, and Yun-Jung Choi for identifying several typos and errors. Their assistance is much appreciated.